

# Covariant priors and model uncertainty

Giovanni Mana<sup>\*</sup> and Carlo Palmisano<sup>†</sup>

**Abstract.** In the application of Bayesian methods to metrology, pre-data probabilities play a critical role in the estimation of the model uncertainty. Following the observation that distributions form Riemann’s manifolds, methods of differential geometry can be applied to ensure covariant priors and uncertainties independent of parameterization. Paradoxes were found in multi-parameter problems and alternatives were developed; but, when different parameters are of interest, covariance may be lost. This paper overviews information geometry, investigates some key paradoxes, and proposes solutions that preserve covariance.

**MSC 2010 subject classifications:** Primary 62C10, 62F15; secondary 62F03, 62F07, 68T37.

**Keywords:** covariant prior distributions, model uncertainty, hierarchical models, model selection, information geometry, Stein paradox, Neyman-Scott paradox.

## 1 Introduction

In metrology, the Bayes theorem allows the information about a measurand – jointly delivered by the data and prior knowledge – to be summarized by probability assignments to its values [Jaynes and Bretthorst (2003); D’Agostini (2003); MacKay (2003); Sivia and Skilling (2006); von der Linden et al. (2014)]. This synt hesis is an essential step to express the measurement uncertainty and to take decisions based on the measurand value.

Including the model in the assessment of the measurement uncertainty requires model selection and averaging [Dose (2007); Elster and Toman (2010); Mana et al. (2012); Toman et al. (2012); Mana et al. (2014); Mana (2015)]; in turn, this requires the model probabilities, which probabilities are proportional to the marginal likelihood, also termed model evidence. Reporting the posterior distribution – or a summary, like point and interval estimates – and marginal likelihood – to carry out model selection or to estimate the model uncertainty – is part of the data analysis. To make a meaningful posterior distribution and uncertainty assessment, the prior density must be covariant; that is, the prior distributions of different parameterizations must be obtained by transformations of variables. Furthermore, it is necessary that the prior densities are proper.

If a preferred parameterization exists, the prior density of any other parameterization is obtained by transformations of variables. We will not examine how to include explicitly-stated information into prior probability distributions. The idea is to find the

---

<sup>\*</sup>INRIM – Istituto Nazionale di Ricerca Metrologica, str. delle Cacce 91, 10135 Torino, Italy  
[g.mana@inrim.it](mailto:g.mana@inrim.it)

<sup>†</sup>UNITO – Università di Torino, Dipartimento di Fisica, v. P. Giuria 1, 10125 Torino, Italy  
[palmisan@to.inf.it](mailto:palmisan@to.inf.it)

maximum-entropy prior subject to the given constraints; discussions can be found in [Jaynes and Bretthorst \(2003\)](#); [D’Agostini \(2003\)](#); [MacKay \(2003\)](#); [Sivia and Skilling \(2006\)](#); [von der Linden et al. \(2014\)](#). If the only information is the statistical model explaining the data, since parametric models form Riemann’s manifolds, priors derived from the manifold metric – for instance, but not necessarily, the Jeffreys ones – ensure the sought covariance, while finite manifold volumes ensure proper priors. The prior density is a part of the model: non-covariant priors go with different metrics and, thus, with different models. To avoid that metrics affect model selection and uncertainty, the prior determining-rule must ensure the metric invariance, that is, different parametric models explaining the same data must have the same metric.

Differential geometry was introduced in statistic by Rao [[Rao \(1945\)](#)], who triggered the development of what is now known as information geometry [[Amari et al. \(2007\)](#); [Arwini and Dodson \(2008\)](#)]. The same ideas led Jeffreys to determine covariant priors from the information metric [[Jeffreys \(1946, 1998\)](#)]. The sections from 2 to 5 give to metrologists an overview of the methods of information geometry as applied to encode the model information into covariant distributions of the parameters and to overcome the feeling of lack of objectivity when carrying out Bayesian inferences. More extensive reviews are in [Kass \(1989\)](#); [Kass and Wasserman \(1996\)](#); [Costa et al. \(2014\)](#).

When the Jeffreys rule is used in multi-parameter problems, paradoxes and inconsistencies were observed and alternatives were developed [[Bernardo \(1979\)](#); [Kass and Wasserman \(1996\)](#); [Datta and Ghosh \(1995\)](#)]. A milestone are the Berger and Bernardo reference priors, that divide the model parameters in parameters of interest – i.e., the measurands – and nuisance parameters [[Berger and Bernardo \(1992a,b\)](#); [Bernardo \(2005\)](#); [Berger and Sun \(2008\)](#); [Berger et al. \(2009, 2015\)](#); [Bodnar et al. \(2015\)](#)]. Roughly, they maximise the expected information gain – the Kullback-Leibler divergence between the prior and posterior distributions – in the limit of infinite measurement repetitions. For regular models where asymptotic normality of the posterior distribution holds, if there aren’t nuisance parameters, the reference priors are the Jeffreys ones [[Ghosh \(2011\)](#)]. However, if there are nuisance parameters, the reference priors depend on the measurands and covariance is lost [[Datta and Ghosh \(1996\)](#)].

The section 6 investigates some of the key inconsistencies and paradoxes reported in the literature. The aim is not to challenge the existence of the inconsistencies or the Berger and Bernardo reference priors, but to show that solutions that save the prior covariance exist. Firstly, we examine why, when a uniform prior is used to infer the measurands’ values from repeated Gaussian measurements, there is no greater uncertainty with many being estimated than with only one [[Jeffreys \(1946, 1998\)](#)]. Next, since when using covariant priors the expected frequencies depend on the cell number, it considers the problem of determining how many outcomes to include in a multinomial model [[Bernardo \(1989\)](#); [Berger et al. \(2015\)](#)]. Eventually, after recognising the diversity and uncertainty of the underlying models, it proposes covariant solutions to the Stein [[Stein \(1959\)](#); [Attivissimo et al. \(2012\)](#); [Samworth \(2012\)](#); [Carobbi \(2014\)](#)] and Neyman-Scott paradoxes [[Neyman and Scott \(1948\)](#)].

## 2 Background

This section introduces Bayesian data analysis, information geometry, and model selection. It does not aim at giving an exhaustive review, but it is a courtesy to metrologists and readers who do not master these fields.

Let us consider the measurement of a scalar quantity  $\alpha$  and the sampling distribution  $p(x|\alpha)$  of the measurement result  $x$  if the measurand value is  $\alpha$ , where we use the same symbol both to indicate random variables and to label the space of their values. The extension to many measurands, the presence of nuisance parameters, and multivariate distributions does not require new concepts and will not be examined in detail. The distribution family  $\mathcal{M} = \{p(x|\alpha) : \alpha \in \mathbb{R}\}$ , whose elements are parameterized by  $\alpha$ , is the parametric model of the data. It is worth noting that  $\alpha$  is a coordinate labelling the  $\mathcal{M}$ 's elements.

The post-data distribution of the measurand values, which updates the information available prior the measurement and synthesized by the prior density  $\pi(\alpha|\mathcal{M})$ , is

$$p(\alpha|x, \mathcal{M}) = \frac{L(\alpha|x)\pi(\alpha|\mathcal{M})}{Z(x|\mathcal{M})}, \quad (2.1)$$

where  $L(\alpha|x) = p(x|\alpha)$  is the likelihood of the model parameters – which is the sampling distribution itself, now, read as a function of the model parameters given the data. The normalizing constant, which is termed marginal likelihood or evidence,

$$Z(x|\mathcal{M}) = \int_{-\infty}^{+\infty} L(\alpha|x)\pi(\alpha|\mathcal{M}) d\alpha \quad (2.2)$$

is the probability distribution of the data given the model. As such, it must be independent of the model parameters. Section 5 will show that (2.2) is crucial in the evaluation of how much the data support the explanation  $\mathcal{M}$ . The symbols  $\pi(\cdot|\mathcal{M})$  and  $p(\cdot|\cdot)$  will be used to indicate prior and posterior probability densities and the relevant conditioning, not given functions of the arguments.

If no information is available,  $\pi(\alpha|\mathcal{M})$  must not deliver information about  $\alpha$ . When it is a continuous variable, we can consider the continuous limit of  $\pi_n = \text{Prob}(\alpha \in \Delta_n)$ , where  $\Delta_n$  is the length of the  $n$ -th interval in which the measurand domain has been subdivided and  $1 \leq n \leq N$ . Hence,  $\pi(\alpha|\mathcal{M}) = \lim_{N \rightarrow \infty} \pi_n / \Delta_n$ , where  $\max(\Delta_n) \rightarrow 0$ . Since, in the class of the finitely discrete ones, the uninformative distribution is  $\pi_n = 1/N$ , the sought prior density is seemingly found.

However, different limit procedures originate different distributions and nothing indicates what should be preferred. For instance, if  $\Delta_n = A/N$ , where  $A$  is the range of the measurand values,  $\pi(\alpha|\mathcal{M}) = 1/A$  will follow. If  $\Delta_n = 1/[\mu(\alpha)N]$ , where  $\mu(\alpha)$  is any probability measure, the prior density is  $\pi(\alpha|\mathcal{M}) = \mu(\alpha)$ . In the same way, if the measurand is changed to  $\beta(\alpha)$ , the  $\beta$ 's distribution,

$$\pi'(\beta|\mathcal{M}) = \pi[\alpha(\beta)|\mathcal{M}] \left| \frac{d\alpha}{d\beta} \right|, \quad (2.3)$$

where  $\pi(\alpha|\mathcal{M})$  represents the absence of information and  $\alpha(\beta)$  is the inverse transformation, is, in general, a different function. For instance, if  $\beta = \alpha^2$ , the transformed distribution is  $\pi'(\beta|\mathcal{M}) = \pi(\sqrt{\beta}|\mathcal{M})/\sqrt{4\beta}$ . Also in this case, since we are as ignorant about  $\beta$  as about  $\alpha$ , nothing indicates what prior distribution is to be used.

These difficulties arise because we assumed that no information is available. But this is not true: as the conditioning on  $\mathcal{M}$  indicates,  $\alpha$  labels the elements of the model explaining the data. It is a way to get a representation, to write formulae explicitly, but, in principle, its choice is arbitrary. Hence, a measurand change corresponds to a mere coordinate change. A parameter-free way to express ignorance is to require that the elements of  $\mathcal{M}$  are equiprobable. Hence, we must supply a metric; this can be done in a natural way, as it will be presently shown.

### 3 Information geometry

This section overviews the methods of information geometry as applied to encode by covariant rules the information delivered by the data models into prior densities. To reduce the algebra to a minimum, we consider only the single measurand case; full treatments can be found in [Amari et al. (2007); Arwini and Dodson (2008)].

Let us introduce the probability amplitudes  $\{\psi(x|\alpha) = \sqrt{p(x|\alpha)} : p(x|\alpha) \in \mathcal{M}\}$ . Since they are a subset of the  $\mathcal{L}^2$  space of the square-integrable functions,  $\mathcal{M}$  inherits the 2-norm metric

$$D(\alpha_2, \alpha_1) = \int_{-\infty}^{+\infty} |\psi(x|\alpha_2) - \psi(x|\alpha_1)|^2 dx \quad (3.1)$$

induced by the  $\mathcal{L}^2$  scalar product. To obtain the metric tensor, we observe that the line element is

$$ds^2 = D(\alpha + d\alpha, \alpha) = \left( \int_{-\infty}^{+\infty} |\partial_\alpha \psi(x|\alpha)|^2 dx \right) d\alpha^2. \quad (3.2)$$

Next, since

$$\partial_\alpha \psi = \frac{\psi \partial_\alpha \ln(\psi^2)}{2}, \quad (3.3)$$

where the units have been chosen in such a way to make the  $x$  variable dimensionless, we can rewrite (3.2) as

$$\begin{aligned} 4ds^2 &= \left( \int_{-\infty}^{+\infty} |\partial_\alpha \ln[p(x|\alpha)]|^2 p(x|\alpha) dx \right) d\alpha^2 \\ &= \langle |\partial_\alpha \ln(L)|^2 \rangle d\alpha^2 = J(L; \alpha) d\alpha^2, \end{aligned} \quad (3.4)$$

where  $L = L(\alpha|x)$  is the likelihood, the angle brackets indicate the average and  $J(L; \alpha)$  is the Fisher information, which is proportional to the metric tensor. It is worth noting that, if  $\partial_\alpha^2 \ln(L)$  exists, the Fisher information can also be written as

$$J(L; \alpha) = -\langle \partial_\alpha^2 \ln(L) \rangle. \quad (3.5)$$

The metric tensor is related to the Kullback-Leibler divergence,

$$D_{KL}(\alpha_2, \alpha) = \int_{-\infty}^{+\infty} \ln \left[ \frac{p(x|\alpha_2)}{p(x|\alpha)} \right] p(x|\alpha_2) dx, \quad (3.6)$$

as follows. By expanding (3.6) in series of  $d\alpha = \alpha_2 - \alpha$  and taking the normalization of  $p(x|\alpha)$  into account, we obtain

$$D_{KL}(\alpha + d\alpha, \alpha) = \frac{1}{2} J(L; \alpha) d\alpha^2, \quad (3.7)$$

up to higher order terms. Therefore,  $ds^2$  measures the information gain when  $p(x|\alpha + d\alpha)$  updates  $p(x|\alpha)$ .

When there are  $p$  parameters, so that  $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_p]^T$  is a  $p \times 1$  matrix, the Fisher information is the  $p \times p$  matrix  $\mathbb{J}(L; \boldsymbol{\alpha}) = -\langle \nabla_{\boldsymbol{\alpha}}^T \nabla_{\boldsymbol{\alpha}} \ln(L) \rangle = \langle \nabla_{\boldsymbol{\alpha}}^T \ln(L) \nabla_{\boldsymbol{\alpha}} \ln(L) \rangle$ , where  $\nabla_{\boldsymbol{\alpha}} = [\partial_1, \partial_2, \dots, \partial_p]$ , and

$$ds^2 \propto d\boldsymbol{\alpha}^T \mathbb{J}(L; \boldsymbol{\alpha}) d\boldsymbol{\alpha}. \quad (3.8)$$

## 4 Covariant priors

The prior for the model parameters must embed the parameter meaning in the problem at hand, which is defined by the model that is assumed to explain the data [von der Linden et al. (2014)]. Therefore, it makes sense to chose a prior density that is the uniform measure of  $\mathcal{M}$  with respect to its metric, that is,

$$\text{Prob}[p(x|\boldsymbol{\alpha}) \in dV_{\boldsymbol{\alpha}}] \propto dV_{\boldsymbol{\alpha}}, \quad (4.1)$$

where, by using the  $\alpha$  parameterization, the volume element is

$$dV_{\boldsymbol{\alpha}} = \sqrt{\det(\mathbb{J})} d\alpha_1 \dots d\alpha_p. \quad (4.2)$$

This induces the Jeffreys probability density [Jeffreys (1946, 1998)]

$$\pi_J(\boldsymbol{\alpha}|\mathcal{M}) \propto \sqrt{\det(\mathbb{J})}, \quad (4.3)$$

which is the continuous limit of a discrete distribution defined over a lattice where the node spacings ensure that the same information is gained when the sampling distribution labelled by a node updates those identified by its neighbourhoods.

The relevance of the Jeffreys rule in metrology and in expressing uncertainties in measurements resides in the metric invariance – that is, the same metric is assumed for all the models that explain the data – and prior covariance under one-to-one coordinate transformations – that is, under reparameterization of the sampling distribution. Covariance makes the post-data distribution (2.1) consistent with transformations of the model parameters. In fact, the left-hand side of (2.1) transforms according to the usual change-of-variable rule. What happens to the right-hand side is that the transformation Jacobian combines with  $\mathbb{J}(L, \boldsymbol{\alpha})$  to give the Fisher information about the new variables. This is a consequence of the invariance of the volume element (4.2).

Here is the explicit proof in the single-parameter case. Let  $\beta(\alpha)$  be a one-to-one coordinate transformation, i.e., a reparameterization of the sampling distribution. By application of (2.1), the post-data distribution of the  $\beta$  measurand is

$$p'(\beta|x, \mathcal{M}) \propto \frac{L'(\beta|x)\sqrt{J(L'; \beta)}}{Z'(x|\mathcal{M})}, \quad (4.4)$$

where the normalization factor of  $\sqrt{J(L'; \beta)}$  has been omitted,  $L'(\beta|x) = L[\alpha(\beta)|x]$  is the reparameterized likelihood,  $\alpha(\beta)$  is the inverse transformation, and  $\pi'_J(\beta) \propto \sqrt{J(L'; \beta)}$  is the Jeffreys prior of  $\beta$ . Since

$$\begin{aligned} \sqrt{J(L; \alpha)} \left| \frac{d\alpha}{d\beta} \right| &= \sqrt{\left\langle \left| \frac{d\alpha}{d\beta} \partial_\alpha \ln(L) \right|^2 \right\rangle} \\ &= \sqrt{\left\langle |\partial_\beta \ln(L)|^2 \right\rangle} = \sqrt{J(L'; \beta)}, \end{aligned} \quad (4.5)$$

by applying the change of variable rule to the  $p(\alpha|x, \mathcal{M})$  post-data distribution of  $\alpha$  in (2.1), we obtain

$$\begin{aligned} p'(\beta|x, \mathcal{M}) &= p[\alpha(\beta)|x, \mathcal{M}] \left| \frac{d\alpha}{d\beta} \right| \\ &\propto \frac{L[\alpha(\beta)|x] \sqrt{J[L; \alpha(\beta)]} \left| \frac{d\alpha}{d\beta} \right|}{Z(x|\mathcal{M})} \\ &= \frac{L'(\beta|x) \sqrt{J(L'; \beta)}}{Z(x|\mathcal{M})}. \end{aligned} \quad (4.6)$$

Eventually, it is easy to prove that (2.1) and (4.6) deliver the same marginal likelihood. In fact,

$$\begin{aligned} Z(x|\mathcal{M}) &\propto \int_{-\infty}^{+\infty} L(\alpha|x) \sqrt{J(L; \alpha)} d\alpha \\ &= \int_{-\infty}^{+\infty} L'(\beta|x) \sqrt{J(L'; \beta)} d\beta \propto Z'(x|\mathcal{M}), \end{aligned} \quad (4.7)$$

where the same normalization factor – by virtue of (4.5) – of  $\sqrt{J(L; \alpha)}$  and  $\sqrt{J(L'; \beta)}$  has been left out. This identity expresses that the marginal distribution of the data is independent of the representation of the model elements.

## 5 Model selection and uncertainty

Let the data be explained by a set of mutually exclusive models that are hyper-parameterized by  $k$ , which can be both discrete (e.g., the cell label of a multinomial

model) or continuous (e.g., the domain boundary of the mean of a Gaussian models). Assuming that the  $\{\mathcal{M}_k\}$  set is complete, by application of the Bayes theorem, the posterior odds on  $\mathcal{M}_k$  explaining the data are [MacKay (2003); Sivia and Skilling (2006); von der Linden et al. (2014)]

$$\text{Prob}(k|x) = \frac{Z(x|k)\pi(k)}{\sum_k Z(x|k)\pi(k)}, \quad (5.1)$$

where  $\pi(k)$  is the prior probability of  $\mathcal{M}_k$  and, when,  $k$  is continuous, an integration substitutes for the sum. In (5.1), the marginal likelihood of a previous Bayesian analysis,  $Z(x|k)$ , is the model likelihood. It is worth noting that, since it is the sampling distributions of the data given  $\mathcal{M}_k$ ,  $Z(x|k)$  must be independent of the model parameters  $\alpha$ ; in turn, the prior density  $\pi(\alpha|k)$  must be covariant.

The marginal likelihood is proportional to the power of the model to explain the data – the higher is the fitness, the greater  $Z(x|k)$  – but inversely proportional to the volume of the parameter space – the greater is the model freedom, the lesser  $Z(x|k)$ . This characteristic of  $Z(x|k)$ , known as the Ockham’s razor [MacKay (2003); Sivia and Skilling (2006); von der Linden et al. (2014)], penalizes the models that, by adjusting the model parameters, have a greater freedom to explain the data. Hence, improper priors, which always correspond to infinite parameter volumes and freedom, make the probability of observing the data set null and, consequently, are never supported; also if they correspond to proper posterior distributions. This issue will be further examined in the next section.

Ensuring that marginal likelihoods do not depend on the  $\mathcal{M}_k$  parameterization makes it possible to calculate the model uncertainty and, in turn, it makes it possible to include the model uncertainty in the error budget [Clyde and George (2004)]. In the same way as  $p(\alpha|x, k)$  expresses the uncertainty of the measurand value – provided that  $\mathcal{M}_k$  explains the data –  $\text{Prob}(k|x)$  expresses the model uncertainty. By combining the measurand and model distributions and by marginalising over the models, the total uncertainty is expressed by

$$p(\alpha|x) = \sum_k p(\alpha|x, k)\text{Prob}(k|x), \quad (5.2)$$

where, if  $k$  is continuous, an integration substitutes for the sum.

## 6 Paradox analyses

In the following, to avoid increasing too much the length of the paper, intermediate calculations are omitted. All were carried out with the help of Mathematica<sup>®</sup> [Wolfram Research Inc. (2012)].

### 6.1 Gaussian model with standardized mean

To exemplify the issues of the use of non-covariant priors, let us consider  $n$  independent samples  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$  from the Gaussian model  $\mathcal{M}_\mu = \{N(x_i|\mu, \sigma) : (\mu, \sigma) \in$

$\mathbb{R} \times \mathbb{R}^+$ , where both the mean  $\mu$  and standard deviation  $\sigma$  are unknown. To simplify the analysis, we set the data offset and measurement unit in such a way that the sample mean and (biased) standard deviation are zero and one, respectively. Hence, the sampling distribution of  $\mathbf{x}$  is

$$p(\mathbf{x}|\mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{n(1+\mu^2)}{2\sigma^2}\right]. \quad (6.1)$$

The Berger and Bernardo reference prior [Bernardo (1979); Berger et al. (2015)] to make inferences about the  $\{\mu, \sigma\}$  pair is

$$\pi_{\text{BB}}(\mu, \sigma|\mathcal{M}_\mu) \propto 1/\sigma, \quad (6.2)$$

which results in the marginal likelihood

$$Z(\mathbf{x}|\mathcal{M}_\mu) = \int_0^\infty \int_{-\infty}^{+\infty} p(\mathbf{x}|\mu, \sigma) \pi_{\text{BB}}(\mu, \sigma|\mathcal{M}_\mu) d\mu d\sigma \propto \frac{\Gamma(n/2 - 1/2)}{2\sqrt{n^n \pi^{n-1}}}, \quad (6.3)$$

where  $\Gamma(z)$  is the Euler gamma function, and posterior distribution

$$p(\mu, \sigma|\mathbf{x}, \mathcal{M}_\mu) = \frac{\sqrt{n^n}}{\sqrt{2^{n-2}\pi} \sigma^{n+1} \Gamma(n/2 - 1/2)} \exp\left[-\frac{n(1+\mu^2)}{2\sigma^2}\right]. \quad (6.4)$$

The proportionality sign in (6.3) stresses that, since  $\pi_{\text{BB}}(\mu, \sigma|\mathcal{M}_\mu)$  is improper; hence, the marginal likelihood is defined only up to an undefined factor. Actually, the normalizing factor of  $\pi_{\text{BB}}(\mu, \sigma|\mathcal{M}_\mu)$  is infinite; therefore,  $Z(\mathbf{x}|\mathcal{M}_\mu)$  is null. This means that the prior (6.2), more precisely its support, is falsified by the data. This problem will be carefully analysed in the following.

If the measurands are the standardized mean  $\lambda = \mu/\sigma$  and  $\sigma$ , the data model is re-parameterized as  $\mathcal{M}_\lambda = \{N(x_i|\sigma\lambda, \sigma) : (\lambda, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$ . Hence, the sampling distribution (6.1) is reparameterized as

$$p(\mathbf{x}|\lambda, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{n(1+\lambda^2\sigma^2)}{2\sigma^2}\right]. \quad (6.5)$$

The Berger and Bernardo reference prior [Bernardo (1979); Berger et al. (2015)] to make inferences about the  $\{\lambda, \sigma\}$  pair is

$$\pi_{\text{BB}}(\lambda, \sigma|\mathcal{M}_\lambda) \propto \frac{1}{\sqrt{2+\lambda^2}\sigma}, \quad (6.6)$$

which results in the marginal likelihood [Wolfram Research Inc. (2012)]

$$\begin{aligned} Z(\mathbf{x}|\mathcal{M}_\lambda) &= \int_0^\infty \int_{-\infty}^{+\infty} p(\mathbf{x}|\lambda, \sigma) \pi_{\text{BB}}(\lambda, \sigma|\mathcal{M}_\lambda) d\lambda d\sigma \\ &\propto \frac{K_0(n/2) \Gamma(n/2 + 1) e^{n/2}}{n \sqrt{(n\pi)^n}}, \end{aligned} \quad (6.7)$$



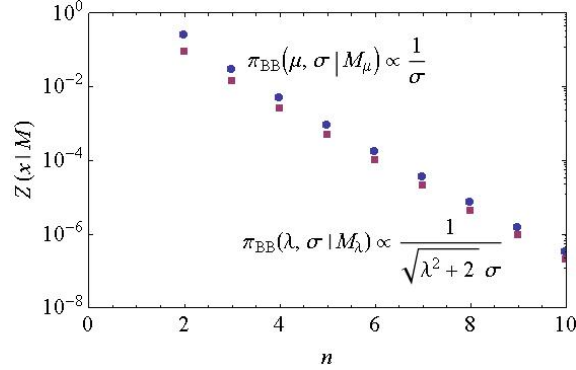


Figure 1: Marginal likelihoods of  $n$  independent data drawn from the same Gaussian distribution and the Berger and Bernardo reference priors (6.2), blue, and (6.6), magenta.

where  $K_0(z)$  is the modified Bessel function of the second kind, and posterior distribution

$$p(\lambda, \sigma | \mathbf{x}, \mathcal{M}_\lambda) = \frac{n\sqrt{n^n}}{\sqrt{2^n(2 + \lambda^2)} \sigma^{n+1} K_0(n/2) \Gamma(n/2 + 1)} \times \exp \left[ -\frac{n(1 + (1 + \lambda)\sigma^2)}{2\sigma^2} \right]. \quad (6.8)$$

Also in this case, since (6.6) is improper, the proportionality sign in (6.7) stresses that  $Z(\mathbf{x} | \mathcal{M}_\lambda)$  is defined only up to an undefined factor; actually, in the limit when the prior support is  $\mathbb{R} \times \mathbb{R}^+$  it is zero.

The priors (6.2) and (6.6) are tailored to the different measurands, but they are not covariant. This originates a twofold difficulty. Firstly, notwithstanding they are conditioned to the same sampling distribution, if we insist on such an interpretation, the data probabilities derived from (6.3) and (6.7) are different, as also shown in Fig. 1. Secondly, after having the distribution (6.4), one can legitimately use it to calculate  $p(\lambda, \sigma | \mathbf{x}, \mathcal{M}_\lambda)$  by changing the variables from  $\{\mu, \sigma\}$  to  $\{\lambda, \sigma\}$ . Also in this case, notwithstanding it is conditioned to the same data, sampling distribution, and prior information, the result is different from (6.8).

The paradox is explained by observing that the prior densities (6.2) and (6.6) embed different information. Since they have different metrics,  $\mathcal{M}_\lambda$  is not a re-parameterization of  $\mathcal{M}_\mu$  and, therefore, the two analyses rely on different models. Bayesian model selection might be used to choose between the competing models. However, one must be aware that, in this case, the selection will act on the metrics of the sampling distributions. A second issue is that, since (6.2) and (6.6) are improper, the marginal likelihood (6.3) and (6.7) are defined only apart undefined proportionality factors. This makes model selection meaningless.

## 6.2 Multinormal distribution

Jeffreys [Jeffreys (1946, 1998)] modified the rule (4.3) because, as we can understand, when applied to independent Gaussian measurements of many measurands having common variance, the degrees of freedom of the posterior distributions of any measurand subset, depend only on the number of measurements, regardless of the measurand number. This implies that the variance of any measurand is the same, no matter how many are estimated. In this section, we will examine this issue in detail.

### Problem statement.

Let  $\mathbb{X} = \{x_{ij}\}$  be  $n$  realizations of  $m$  independent normal variables with unknown means  $\mu_i$  and variance  $\sigma^2$ , for  $i = 1, 2, \dots, m$  and  $j = 1, 2, \dots, n$ . The sampling distribution of  $\mathbb{X}$  is

$$p(\mathbb{X}|\boldsymbol{\mu}, \sigma) = \prod_{j=1}^n N_m(\mathbf{x}_j|\boldsymbol{\mu}, \sigma^2 \mathbb{I}_m), \quad (6.9)$$

where  $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_m]^T$ ,  $\mathbf{x}_j = [x_{1j}, x_{2j}, \dots, x_{mj}]^T$ , and  $\mathbb{I}_m$  is the  $m \times m$  identity matrix. The Jeffreys prior of  $\{\boldsymbol{\mu}, \sigma\}$  is

$$\pi_J(\boldsymbol{\mu}, \sigma|\mathcal{M}) = \frac{m\sigma_0^m}{V_\mu \sigma^{m+1}} \quad (6.10)$$

and  $\mathcal{M}$  is the data model where  $V_\mu$  is the volume of the  $\boldsymbol{\mu}$  subspace and  $\sigma > \sigma_0$ . In the following, the (6.10) support will be identified with  $\mathbb{R}^m \times \mathbb{R}^+$ . However, it is unknown and should be chosen by model selection. This problem will be examined in section 6.4.

The joint post-data distribution of  $\boldsymbol{\mu}$  and  $\sigma$  is

$$\begin{aligned} p(\boldsymbol{\mu}, \sigma|\mathbb{X}, \mathcal{M}) &= \frac{p(\mathbb{X}|\boldsymbol{\mu}, \sigma)}{Z(\mathbb{X}|\mathcal{M})} \frac{m\sigma_0^m}{V_\mu \sigma^{m+1}} \\ &= \frac{1}{Z(\mathbb{X}|\mathcal{M})} \frac{\exp\left(-\frac{mn\bar{s}^2}{2\sigma^2}\right) \exp\left(-\frac{n|\bar{\mathbf{x}} - \boldsymbol{\mu}|^2}{2\sigma^2}\right)}{\sqrt{(2\pi)^{mn}} \sigma^{mn}} \frac{m\sigma_0^m}{V_\mu \sigma^{m+1}}, \end{aligned} \quad (6.11)$$

where, if the integration domain is extended to  $\mathbb{R}^m \times \mathbb{R}^+$ ,

$$\begin{aligned} Z(\mathbb{X}|\mathcal{M}) &= \int_{\sigma_0}^{\infty} \int_{V_\mu} p(\mathbb{X}|\boldsymbol{\mu}, \sigma) \pi_J(\boldsymbol{\mu}, \sigma|\mathcal{M}) d\boldsymbol{\mu} d\sigma \\ &\approx \frac{\sqrt{(2m)^m} \Gamma(mn/2)}{2\sqrt{(mn)^{m(n+1)}} \pi^{m(n-1)} \bar{s}^{mn}} \frac{m\sigma_0^m}{V_\mu} \end{aligned} \quad (6.12)$$

is the marginal likelihood,  $\Gamma(z)$  is the Euler gamma function,  $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m]^T$ ,  $\bar{x}_i$  and  $\bar{s}_i^2$  are the mean and (biased) variance of the data set  $\{x_{i1}, x_{i2}, \dots, x_{in}\}$ , and  $\bar{s}^2 = (1/m) \sum_{i=1}^m \bar{s}_i^2$  is the pooled variance. Although the  $m\sigma_0^m/V_\mu$  factor of (6.12)

cancels and (6.11) is proper, when  $\sigma_0 \rightarrow 0$  or  $V_\mu \rightarrow \infty$ , the data probability  $Z(\mathbb{X}|\mathcal{M})$  is null.

The joint post-data distribution of any subset  $\boldsymbol{\mu}' = [\mu_1, \mu_2, \dots, \mu_q]^T$  of  $q \leq m$  means is the  $q$ -variate Student distribution having location  $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_q]^T$ , scale matrix  $\bar{s}^2 \mathbb{I}_q/n$ , and  $mn$  degrees of freedom

$$\begin{aligned} t_{mn}(\boldsymbol{\mu}'|\mathbb{X}, \mathcal{M}) &= \int_{V_\mu} \int_{\sigma_0}^{\infty} p(\boldsymbol{\mu}, \sigma|\mathbb{X}, \mathcal{M}) d\sigma d\mu_{q+1} \dots d\mu_m \\ &\approx \frac{\Gamma[(mn+q)/2](1+t^2)^{-(mn+q)/2}}{\sqrt{(m\pi)^q} \Gamma(mn/2) \bar{s}^q}, \end{aligned} \quad (6.13)$$

where integration domain is again extended to  $\mathbb{R}^m \times \mathbb{R}^+$  and  $t^2 = |\bar{\mathbf{x}}' - \boldsymbol{\mu}'|^2 / (m\bar{s}^2)$ . The mean and variance-covariance matrix of  $\boldsymbol{\mu}'$  are

$$E(\boldsymbol{\mu}'|\mathbb{X}, \mathcal{M}) = \bar{\mathbf{x}}' \quad (6.14a)$$

and

$$\Sigma^2(\boldsymbol{\mu}'|\mathbb{X}, \mathcal{M}) = \frac{m\bar{s}^2}{mn-2} \mathbb{I}_q, \quad (6.14b)$$

where  $nm > 2$  and  $\mathbb{I}_q$  is the  $q \times q$  identity matrix.

If  $m = q = 1$ ,  $t = (\mu - \bar{x})/\bar{s}$  is a Student variable having  $n$  degrees of freedom. However, from a frequentist viewpoint, it is a Student variable having  $n - 1$  degrees of freedom. Furthermore, the variance-covariance matrix (6.14b) is independent of the number  $q$  of estimated means. This implies that, for given sample means and pooled variance from  $mn$  observations, there would be no greater uncertainty with  $q$  means being estimated than with only one. In particular, the variance of any measurand is  $\sigma_\mu^2 = m\bar{s}^2/(mn-2)$ , no matter how many are estimated.

### Proposed solution

The degrees of freedom discrepancy is explained by observing that the Bayesian posterior (6.13) assigns probabilities to the winning elements of a measurands population associated with the same  $\mathbb{X}$  sample. Contrary, the frequentist distribution assigns probabilities to the  $t$ -statistic of different  $\mathbb{X}$  samples given the same measurands. Therefore, there is no reason to expect that the two distributions are the same.

The independence of (6.14b) from  $q$  does not mean that the uncertainty is independent of the measurand number. In fact, the one standard-deviation credible region of  $\boldsymbol{\mu}'$  is a  $q$ -ball having  $\sigma_\mu$  radius. The probability that  $\boldsymbol{\mu}'$  is in this region is [Wolfram Research Inc. (2012)]

$$\begin{aligned} \int_{\Omega} \int_0^{\sigma_\mu} t^{q-1} t_{mn}(\boldsymbol{\mu}'|\mathbb{X}) dt d\Omega = \\ \frac{\Gamma[(mn+q)/2]}{\sqrt{(mn-2)^q} \Gamma(mn/2)} {}_2F_1[q/2, (mn+q)/2, (q+2)/2, 1/(2-mn)], \end{aligned} \quad (6.15)$$

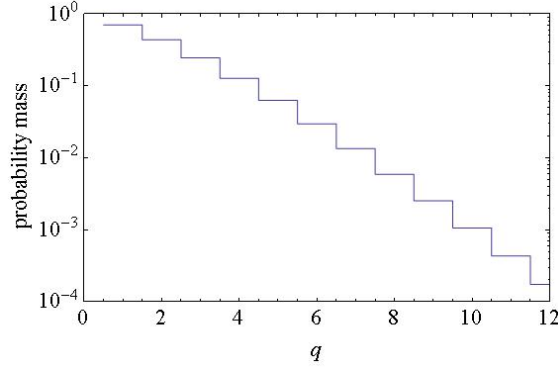


Figure 2: Probability that a  $q$ -subsets of  $m$  measurands is in the one standard-deviation credible region. The data are  $n$  independent and identically Gaussian measures of each measurand. It is worth noting that the probability depends only on  $q$  and the  $mn$  degrees of freedom, which were set to 12.

where  $d\Omega$  is for the  $m - 1$  angular factors and  ${}_2F_1(a, b, c; z)$  is the hypergeometric function. As shown in Fig. 2, as the number of measurands increases, the probability decreases from the maximum – the probability that  $\mu_i$  is in the  $[-\sigma_\mu, +\sigma_\mu]$  interval – to zero. Therefore, despite the variance is the same, there is a greater uncertainty with  $q$  measurands being simultaneously estimated than with only one.

### 6.3 Multinomial paradox

We now consider the problem of determining how many outcomes to include in a multinomial model. When using the Jeffreys rule, the expected outcome frequencies depend on the number of cells. In particular, frequencies depend on the number of void cells and tend to zero as they tend to the infinity. Since one has the option of adding outcomes, a paradox arises when the observations are explained by models including an arbitrary number of cells, where no event is observed [Bernardo (1989); Berger et al. (2015)].

#### Problem statement

Suppose  $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ , where  $x_i$  are positive integers, is a realization of multinomial variable. Hence,  $\mathbf{x} \sim \text{Mu}(\mathbf{x}|m, \boldsymbol{\theta})$ , where  $m$  is the cell number,  $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_m\}$  are frequency-of-occurrence,  $0 \leq \theta_i \leq 1$ ,  $\sum_{i=1}^m \theta_i = 1$ ,

$$\text{Mu}(\mathbf{x}|m, \boldsymbol{\theta}) = \frac{n! \prod_{i=1}^m \theta_i^{x_i}}{\prod_{i=1}^m x_i!}, \quad (6.16)$$

and  $n = \sum_{i=1}^m x_i$  is the sample size. The Jeffreys' prior of  $\boldsymbol{\theta}$  is the Dirichlet distribution

$$\pi_J(\boldsymbol{\theta}|m) = \frac{\Gamma(m/2)}{\sqrt{\pi^m} \prod_{i=1}^m \sqrt{\theta_i}}, \quad (6.17)$$

where  $\Gamma(z)$  is the Euler gamma function. The posterior distribution of  $\boldsymbol{\theta}$  given the counts  $x_i$  and the  $m$ -cell explaining model is

$$p(\boldsymbol{\theta}|\mathbf{x}, m) = \frac{\text{Mu}(\mathbf{x}|m, \boldsymbol{\theta})\pi_J(\boldsymbol{\theta}|m)}{Z(\mathbf{x}|m)} = \frac{\Gamma(n + m/2) \prod_{i=1}^m \theta_i^{x_i-1/2}}{\prod_{i=1}^m \Gamma(x_i + 1/2)}, \quad (6.18)$$

where [Wolfram Research Inc. (2012)]

$$Z(\mathbf{x}|m) = \int_{\boldsymbol{\theta} \in \Theta} \text{Mu}(\mathbf{x}|m, \boldsymbol{\theta})\pi_J(\boldsymbol{\theta}|m) d\boldsymbol{\theta} = \frac{n!\Gamma(m/2) \prod_{i=1}^m \Gamma(x_i + 1/2)}{\sqrt{\pi^m} \Gamma(n + m/2) \prod_{i=1}^m x_i!} \quad (6.19)$$

is the marginal likelihood and  $\Theta$  is the  $m$ -dimensional simplex  $\sum_{i=1}^m \theta_i = 1$ . The posterior mean of  $\theta_i$  is [Bernardo (1989)]

$$E(\theta_i|\mathbf{x}, m) = \frac{x_i + 1/2}{n + m/2}. \quad (6.20)$$

Since (6.20) depends on the cell number, a paradox arises when the counts are explained by models including additional cells, where no event is observed. In particular, (6.20) depends on the number of void-cells and tends to zero as  $m$  tends to the infinity.

### Proposed solution.

If the number of cells is unknown, the model uncertainty must be included into the analysis, as the conditioning over  $m$  indicates. Therefore, (6.20) must be averaged over the models, the average being weighed by the model probabilities

$$\text{Prob}(m|\mathbf{x}) = \frac{Z(\mathbf{x}|m)}{\sum_{l=m_1}^{m_2} Z(\mathbf{x}|l)}, \quad (6.21)$$

where  $m_1$  is the number of non-null elements of  $\mathbf{x}$ ,  $m_2$  is the upper bound to the cell number, and the models are assumed mutually exclusive and equiprobable. The asymptotic behaviour of  $\text{Prob}(m|\mathbf{x})$  is  $1/m^n$ ; hence, models having increasing number of voids cells are less and less probable and contribute less and less to the posterior mean. An example is shown in Fig. 3.

## 6.4 Stein paradox

When estimating the mean power of a number of signal from Gaussian measurements of their amplitudes, uniform priors of the unknown amplitudes lead to problematic posterior power distribution and expectation. In particular, as the number of signals tends to the infinity, the expected posterior power is inconsistent [Stein (1959); Bernardo (1979, 2005); Attivissimo et al. (2012); Samworth (2012); Carobbi (2014); Berger et al. (2015)].

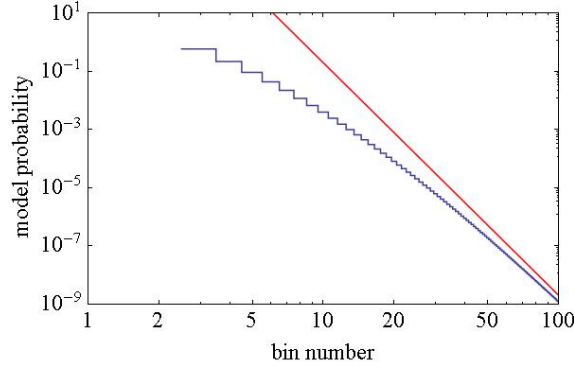


Figure 3: Probability that the multinomial distribution  $\text{Mu}(x_1, x_2, \dots, x_m | \theta_1, \theta_2, \dots, \theta_m)$  for  $m = 3, 4, \dots, 100$  explains the data set  $\mathbf{x} = \{x_1 = 2, x_2 = 1, x_3 = 5, x_4 = 0, \dots, x_m = 0\}$ , where  $n = 8$ . The red line is the  $1/m^8$  asymptote.

### Problem statement.

In the simplest form of the Stein paradox,  $x_i$  are realizations of independent normal variables with unknown means  $\mu_i$  and known variance  $\sigma^2$ , for  $i = 1, 2, \dots, m$ . To keep the algebra simple, without loss of generality, we use units where  $\sigma = 1$ ; when necessary for the sake of clarity, we will write explicitly quantity ratios *vs.*  $\sigma$ . The Jeffreys' prior for each  $\mu_i$  is a constant, resulting in independent Gaussian posteriors,  $N(\mu_i | x_i, \sigma = 1)$ , for each individual mean.

Let us consider  $\mu_i^2$  – from a Bayesian viewpoint, they are independent non-central  $\chi_1^2$  variables having one degree of freedom, mean  $1 + x_i^2$ , and variance  $2(1 + 2x_i^2)$  – and the  $\theta^2 = (\sum_{i=1}^m \mu_i^2)/m$  measurand – from a Bayesian viewpoint, it is a non-central  $\chi_m^2$  variable having  $m$  degrees of freedom. The Bayesian posterior mean and variance of  $\theta^2$  are

$$\overline{\theta^2} = E(\theta^2 | \mathbf{x}, \sigma = 1, \mathcal{M}_\infty) = 1 + \overline{x^2}, \quad (6.22a)$$

where  $\overline{x^2} = \sum_{i=1}^m x_i^2 / m$ , and

$$\sigma_{\theta^2}^2 = \text{Var}(\theta^2 | \mathbf{x}, \sigma = 1, \mathcal{M}_\infty) = \frac{2}{m} (1 + 2\overline{x^2}), \quad (6.22b)$$

where  $\mathbf{x}$  is the  $\{x_1, x_2, \dots, x_m\}$  list and  $\mathcal{M}_\infty$  is the model where the  $\mu_i$  domain is  $\mathbb{R}$ .

From a frequentist viewpoint, since  $\sum_{i=1}^m x_i^2$  is a non-noncentral  $\chi_m^2$  variable having  $m$  degrees of freedom,  $\overline{\theta^2}$  is a biased estimator of  $\theta^2$ . In fact,

$$E(\overline{\theta^2} | \boldsymbol{\mu}, \sigma = 1) = 2 + \theta^2 \quad (6.23)$$

where  $\boldsymbol{\mu}$  is the  $\{\mu_1, \mu_2, \dots, \mu_m\}$  list. The frequentist variance of  $\overline{\theta^2}$  is given by the same (6.22b). As  $m$  tends to infinity a worst situation occurs: (6.22a) and (6.22b) predict that  $\theta^2$  is certainly equal to  $\overline{x^2} + 1$ , but, at the same time, (6.23) and (6.22b) predict that  $\theta^2$  is certainly equal to  $\overline{x^2} - 1$ .

**Proposed solution.**

To explain the paradox, we observe that it occurs only if the  $\mu_i^2/m$  series converges. Otherwise,  $\theta^2 \rightarrow \infty$  and (6.22a) and (6.23) give the same result for all practical purposes. The improper prior  $\pi_J(\mu_i) \propto \text{const.}$  encodes that  $|\mu_i|$  is greater than any positive number. This information is irrelevant to the  $\mu_i$  posterior – besides, the odds on  $\mu_i$  positive or negative are the same; but, it is not so for the  $\mu_i^2$  posterior. If  $|\mu_i| < \infty$ , the  $\mu_i$  domain must be bounded. Furthermore, if  $\theta^2$  exists, these domains must be bounded also when  $m \rightarrow \infty$ .

A statistical model encoding this information is  $\mathcal{M}_{ba} = \{N(\mathbf{x}|\boldsymbol{\mu}, \sigma = 1) : \boldsymbol{\mu} \in [b - \sqrt{3}a, b + \sqrt{3}a]^m\}$ , where  $\mathbf{x}$  is the  $\{x_1, x_2, \dots, x_m\}$  list. In this way the data model gets a boundary and different hyper-parameters correspond to different models. The Jeffreys distribution of  $\mu_i$  is the gate function  $\pi_J(\mu_i|a, b) = U(\mu_i; b - \sqrt{3}a, b + \sqrt{3}a)$ , which is zero outside the  $[b - \sqrt{3}a, b + \sqrt{3}a]$  interval and  $1/(2a)$  inside, the  $a, b$  hyper-parameters being unknown. To infer (6.22a) and (6.22b), the  $a$  parameter was assumed large enough to identify the data model with  $\mathcal{M}_\infty$ . However, this model is not necessarily supported by the data; therefore, model selection and averaging are necessary to take this missing information into account.

To make analytical computations possible, we substitute a Gaussian density for the  $U(\mu_i; b - \sqrt{3}a, b + \sqrt{3}a)$  prior. We have not yet verified that the inferences made by using a Gaussian prior are not different from those made by using a gate prior. We do not expect qualitative differences, but, if so, it would be interesting to understand why. Hence, in the case of  $m$  observations  $\{x_i\} \sim \prod_{i=1}^m N(x_i|\mu_i, \sigma = 1)$ , the Gaussian approximation of the Jeffreys pre-data distribution of  $\boldsymbol{\mu}$  is

$$\pi_J(\boldsymbol{\mu}|a, b) = \prod_{i=1}^m N(\mu_i|b, a), \quad (6.24)$$

which results in independent and identically distributed  $\mu_i$  having marginal likelihood

$$\begin{aligned} Z(x_i|b, a) &= \int_{-\infty}^{+\infty} N(x_i|\mu_i, \sigma = 1) N(\mu_i|b, a) d\mu_i \\ &= \frac{1}{\sqrt{2\pi(1+a^2)}} \exp\left[-\frac{(x_i - b)^2}{2(1+a^2)}\right] \end{aligned} \quad (6.25)$$

and normal posterior

$$p(\mu_i|x_i, \sigma = 1, b, a) = N(\mu_i|\bar{\mu}_i, \sigma_\mu), \quad (6.26a)$$

having

$$\bar{\mu}_i = E(\mu_i|x_i, \sigma = 1, b, a) = \frac{a^2 x_i + b}{1 + a^2} \quad (6.26b)$$

mean and

$$\sigma_\mu^2 = \text{Var}(\mu_i|\sigma = 1, b, a) = \frac{a^2}{1 + a^2} \quad (6.26c)$$

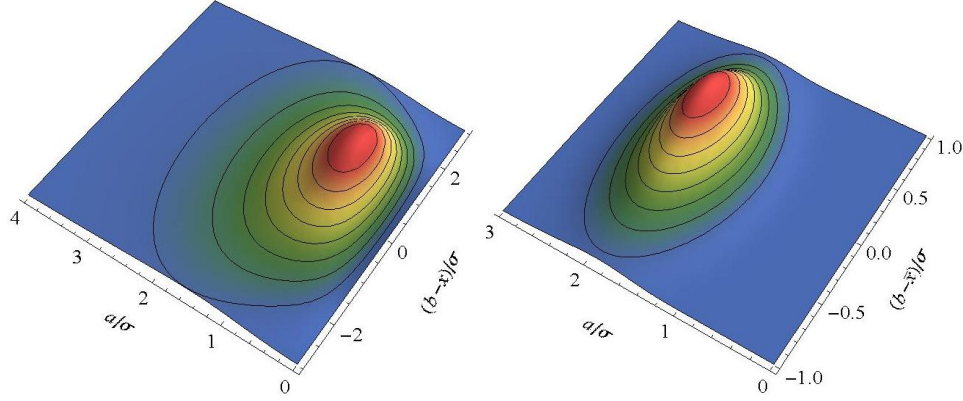


Figure 4: Posterior probability densities of the model  $\mathcal{M}_{ba}$  given the  $\{x_1, x_2, \dots, x_m\}$  Gaussian data. Left:  $m = 1$ . Right:  $m = 20$  and  $s_x/\sigma = 2$

variance. The equation (6.25) benefits of the factorization

$$N(x_i|\mu_i, \sigma)N(\mu_i|b, a) = N(x_i|b, \sqrt{a^2 + \sigma^2})N\left(\mu_i \left| \frac{ax_i^2 + b\sigma^2}{a^2 + \sigma^2}, \frac{a\sigma}{\sqrt{a^2 + \sigma^2}} \right.\right), \quad (6.27)$$

which solves the integration.

By applying the Jeffreys rule to the marginal likelihood,

$$Z(\mathbf{x}|b, a) = \prod_{i=1}^m Z(x_i|b, a) = \frac{\exp\left\{-\frac{m[s_x^2 + (\bar{x} - b)^2]}{2(1 + a^2)}\right\}}{\sqrt{(2\pi)^m(1 + a^2)^m}}, \quad (6.28)$$

which is the probability density of the data given the model and where  $\bar{x} = \sum_{i=1}^m x_i/m$  and  $s_x^2 = \sum_{i=1}^m (x_i - \bar{x})^2/m$  are the sample mean and variance, we obtain the hyper-parameter prior

$$\pi_J(b, a) \propto \frac{a}{\sqrt{(1 + a^2)^3}}. \quad (6.29)$$

Hence, the odds on  $\mathcal{M}_{ba}$  explaining the data are [Wolfram Research Inc. (2012)]

$$\begin{aligned} p(b, a|\bar{x}, s_x^2) &= \frac{Z(\mathbf{x}|b, a)\pi_J(b, a)}{\int_0^\infty \int_{-\infty}^{+\infty} Z(\mathbf{x}|b, a)\pi_J(b, a) db da} \\ &= \frac{\sqrt{2m(ms_x^2/2)^m} a}{\sqrt{\pi(1 + a^2)^{m+3}} \Gamma(m/2, 0, ms_x^2/2)} \exp\left\{-\frac{m[s_x^2 + (b - \bar{x})^2]}{2(1 + a^2)}\right\}, \end{aligned} \quad (6.30)$$

where  $\Gamma(a, z_1, z_2)$  is the generalized incomplete gamma function. Figure 4 shows the probability density (6.30) when  $m = 1$  (left) and  $m = 20$  and  $s_x/\sigma = 2$  (right). It is



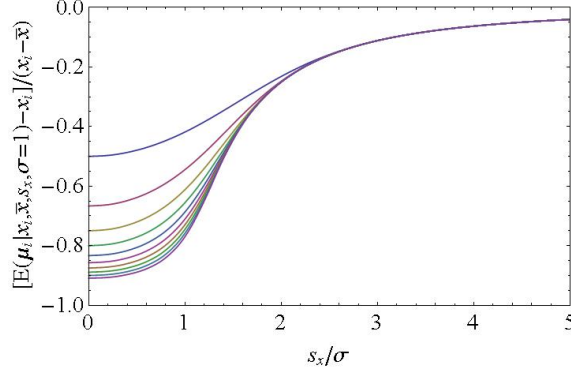


Figure 5: Model-averaged mean of  $\mu_i$  vs. the sample standard-deviation  $s_x$ . Ten cases were considered; from  $m = 2$  (top line) to  $m = 20$  (bottom line), in  $m = 2$  steps.

worth noting that  $p(b, a | \bar{x}, s_x^2)$  depends only on the sample mean and variance and that its maximum occurs at  $b = \bar{x}$ .

Let us discuss the  $m = 1$  case,

$$p(b, a | x) = \frac{a}{\sqrt{2\pi}(1+a^2)^2} \exp \left[ -\frac{(b-x)^2}{2(1+a^2)} \right], \quad (6.31)$$

in some details. Firstly, we observe that not all the  $\mu$  domains are equally supported by the  $x$  datum, in particular the  $\mathbb{R}$  domain is excluded; the model most supported by the data, i.e., the mode of (6.31), is  $\mathcal{M}(x, 1/\sqrt{3})$ . In the second place, after averaging (6.26a) over (6.31), the posterior distribution of  $\mu$  and  $\mu^2$  are

$$p(\mu | x, \sigma = 1) = \int_0^\infty \int_{-\infty}^{+\infty} N(\mu | \bar{\mu}, \sigma_\mu) p(b, a | x) db da = N(\mu | x, \sigma = 1), \quad (6.32a)$$

and the non-central  $\chi_1^2$  distribution,

$$p(\mu^2 | x, \sigma = 1) = \int_{-\infty}^{+\infty} \delta(\mu^2 - \tau^2) N(\tau | x, \sigma = 1) d\tau = \chi_1^2(\mu^2 | x^2), \quad (6.32b)$$

having one degree of freedom and non-centrality parameter  $x^2$ . This is an important and non-trivial result, which demonstrates that the hierarchical model  $\mathcal{M}_{ba}$  is consistent with the one-level model that uses the improper prior  $\pi(\mu | \mathcal{M}_\infty) \propto 1$ . Therefore, there is no hyper-prior effect on the posterior distributions of  $\mu$  and  $\mu^2$ , whose expected values are  $E(\mu | x, \sigma = 1) = x$  and  $E(\mu^2 | x, \sigma = 1) = 1 + x^2$ . However, the  $\mu^2$  expectation conditioned to the mode of the (6.31) distribution – that is, to the  $\mathcal{M}(x, 1/\sqrt{3})$  model most supported by the data – decreases to

$$E(\mu^2 | x, \sigma = 1, b = x, a = 1/\sqrt{3}) = 1/4 + x^2. \quad (6.33)$$

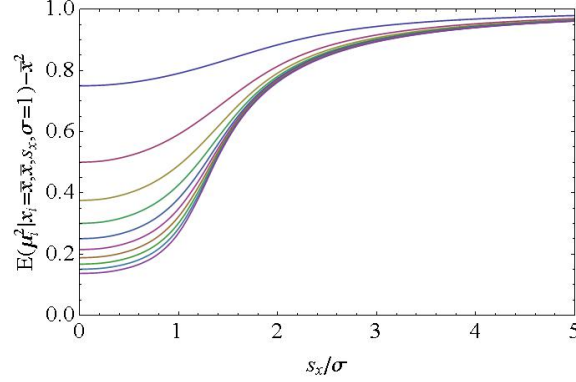


Figure 6: Model-averaged mean of  $\mu_i^2$  vs. the sample standard-deviation  $s_x$ , when  $x_i = \bar{x}$ . Ten cases were considered; from  $m = 2$  (top line) to  $m = 20$  (bottom line), in  $m = 2$  steps.

After averaging the means (6.26b) over the models, we obtain [Wolfram Research Inc. (2012)]

$$\begin{aligned} E(\mu_i | x_i, \bar{x}, s_x, \sigma = 1) &= \int_0^\infty \int_{-\infty}^{+\infty} E(\mu_i | x_i, \sigma = 1, b, a) p(b, a | \bar{x}, s_x^2) db da \\ &= x_i + \frac{2(x_i - \bar{x}) [(m+2)\Gamma(m/2, 0, ms_x^2/2) - 2\Gamma(m/2+2)]}{m(m+2)s_x^2\Gamma(m/2, 0, ms_x^2/2)}. \end{aligned} \quad (6.34)$$

Figure 5 shows the difference between the model-averaged mean and  $x_i$ . When the sample variance  $s_x^2/\sigma^2$  is small,  $E(\mu_i | x_i, \bar{x}, s_x, \sigma = 1) \rightarrow x_i - m(x_i - \bar{x})/(m+2)$  and the model average shrinks towards  $\bar{x}$ ; more are the measurands, the stronger the shrink. Actually, when  $m \rightarrow \infty$ , data consistency requires that the lower bound of the sample variance is  $s_x^2/\sigma^2 = 1$ ; in this case the model average converges to the sample mean. When the sample variance is large,  $E(\mu_i | x_i, \bar{x}, s_x, \sigma = 1) \rightarrow x_i$  and the model average supports the  $x_i$  datum. This is consistent with a large (small) sample variance supporting the hypothesis that the measurands are different (the same).

The  $(\mu_i/\sigma_\mu)^2$  measurands are independent and identical non-central  $\chi_1^2$  variables having one degree of freedom and non-centrality parameter  $\lambda_i^2 = (\bar{\mu}_i/\sigma_\mu)^2$ . Hence,

$$p(\mu_i^2 | x_i, \sigma = 1, b, a) = \frac{\chi_1^2(\mu_i^2/\sigma_\mu^2 | \lambda_i^2)}{\sigma_\mu^2}, \quad (6.35)$$

where  $\bar{\mu}_i$  and  $\sigma_\mu^2$  are given by (6.26b) and (6.26c). The expected  $\mu_i^2$  value is

$$E(\mu_i^2 | x_i, \sigma = 1, b, a) = \sigma_\mu^2 + \bar{\mu}_i^2 = \frac{b^2 + a^2(1 + 2bx_i) + a^4(1 + x_i^2)}{(1 + a^2)^2} \quad (6.36)$$

and, after averaging over the models, we obtain [Wolfram Research Inc. (2012)]

$$\begin{aligned} E(\mu_i^2|x_i, \bar{x}, s_x, \sigma = 1) &= \int_0^\infty \int_{-\infty}^{+\infty} E(\mu_i^2|x_i, \sigma = 1, b, a) p(b, a|\bar{x}, s_x^2) db da \\ &= x_i^2 + 1 - \frac{A \exp(-ms_x^2/2) + B[2\Gamma(m/2 + 2) - (m + 2)\Gamma(m/2 + 1, ms_x^2/2)]}{(m + 2)m^2 s_x^4 \Gamma(m/2, 0, ms_x^2/2)}, \end{aligned} \quad (6.37)$$

where

$$A = \frac{(m + 2)\sqrt{(ms_x^2)^{m+2}}(x_i - \bar{x})^2}{\sqrt{2^{m-2}}} \quad (6.38)$$

and

$$B = 2s_x^2[1 - m - 2mx_i(x_i - \bar{x})] + 2(m + 2)(x_i - \bar{x})^2. \quad (6.39)$$

In addition to the sample variance, the model average (6.37) depends on both  $x_i$  and the sample mean; this is a consequence of the prior information that all  $\mu_i$  belong to the same interval. When the sample variance tends to zero and infinity, the model average converges to [Wolfram Research Inc. (2012)]

$$\lim_{s_x \rightarrow 0} E(\mu_i^2|x_i, \bar{x} = x_i, s_x, \sigma = 1) = x_i^2 + 1 - \frac{m - 1}{m + 2}, \quad (6.40)$$

where  $\bar{x} \rightarrow x_i$  because of consistency, and  $x_i^2 + 1$ , respectively. When  $m \rightarrow \infty$ , the sample variance is bounded by  $s_x^2/\sigma^2 > 1$  and [Wolfram Research Inc. (2012)]

$$\lim_{m \rightarrow \infty} E(\mu_i^2|x_i, \bar{x}, s_x = 1, \sigma = 1) = \bar{x}^2. \quad (6.41)$$

These results are consistent with a large variance indicating different measurands – hence, the model average is  $x_i^2 + 1$  – and a unit variance indicating the same measurand – hence, the model average is  $\bar{x}^2$ . Figure 6 shows the difference between the model average and  $\bar{x}^2$ , when  $x_i = \bar{x}$ . When  $s_x^2/\sigma^2$  is small,  $E(\mu_i^2|x_i, \bar{x}, s_x, \sigma = 1) \rightarrow \bar{x}^2$ ; most are the measurands, the nearest is the average. When the sample variance is big, or there is only one measurand, the model average converges to  $x_i^2 + 1$ .

Eventually, let us turn the attention to  $\theta^2$ . The expected value is

$$E(\theta^2|\bar{x}, s_x, \sigma = 1, b, a) = \frac{\sum_{i=1}^m (\sigma_\mu^2 + \bar{\mu}_i^2)}{m} = \frac{b^2 + a^2(1 + 2b\bar{x}) + a^4(1 + s_x^2 + \bar{x}^2)}{(1 + a^2)^2} \quad (6.42)$$

and, after averaging over the odds on the data models [Wolfram Research Inc. (2012)],

$$\begin{aligned} E(\theta^2|\bar{x}^2, s_x, \sigma = 1) &= \int_0^\infty \int_{-\infty}^{+\infty} E(\theta^2|\bar{x}, s_x, \sigma = 1, b, a) p(b, a|\bar{x}, s_x^2) db da \\ &= 1 + \bar{x}^2 + 2A/B, \end{aligned} \quad (6.43)$$

where

$$A = (3 - 2ms_x^2)\Gamma(m/2 + 1) - 2\Gamma(m/2 + 2, ms_x^2/2) \quad (6.44a)$$

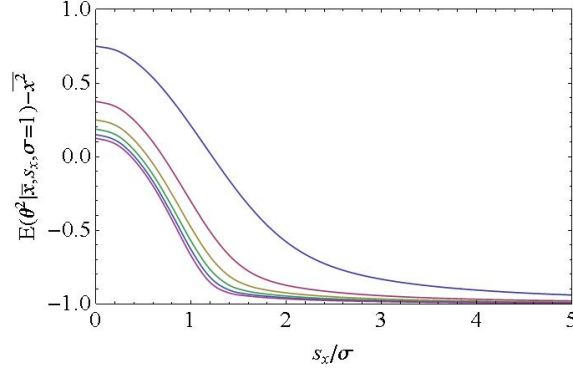


Figure 7: Model-averaged expectation of  $\theta^2$ . Five cases were considered; from  $m = 2$  (top line) to  $m = 22$  (bottom line), in  $m = 4$  steps.

$$-(1 - m - 2ms_x^2)\Gamma(m/2 + 1, ms_x^2/2)$$

and

$$B = m^2 s_x^2 \Gamma(m/2, 0, ms_x^2/2). \quad (6.44b)$$

As shown in Fig. 7, the Stein paradox is solved by observing that, when  $s_x^2/\sigma^2$  tends to zero – that is, when there is a single measurand, the model average converges to the frequentist estimate  $\bar{x}^2$ ; most are the measurands, the nearest are the estimates. In the limit of a big sample-variance – that is, when there are many different measurands, the model average converges to  $\bar{x}^2 - 1$ , which is the frequentist unbiased estimate. It is non-obvious and remarkable that, when  $m \rightarrow \infty$  and, consistently,  $s_x^2/\sigma^2 > 1$ , the model average is always  $\bar{x}^2 - 1$ .

## 6.5 Neyman-Scott paradox

The problem is to estimate the common variance of independent Gaussian observations of different quantities, where the quantity values are not of interest. In the simplest case, this is a fixed effect model with two observations on each quantity. As the number of observed quantities grows without bound, the Jeffreys prior leads to an inconsistent expectation of the common variance [Neyman and Scott (1948)].

### Problem statement.

Let  $\{x_1, x_2\}_i$  be  $m$  independent pairs of independent Gaussian variables having different means  $\mu_i$  and common variance  $\zeta = \sigma^2$ , for  $i = 1, \dots, m$ , all parameters being unknown. By changing the data from  $\{x_1, x_2\}_i$  to the sample means  $\bar{x}_i = (x_{1i} + x_{2i})/2$  and pooled variance  $s^2 = \sum_{i=1}^m s_i^2/m$ , where  $s_i^2 = (x_{1i} - x_{2i})^2/2$ , the sampling distribution is

$$p(\bar{\mathbf{x}}, s^2 | \boldsymbol{\mu}, \zeta) = \frac{2m}{\zeta} \chi_m^2(2ms^2/\zeta) N_m(\bar{\mathbf{x}} | \boldsymbol{\mu}, \zeta \mathbb{I}_m/2), \quad (6.45)$$

where  $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m]^T$ ,  $\chi_m^2(z)$  is a chi-square distribution having  $m$  degrees of freedom, and  $N_m(\bar{\mathbf{x}}|\boldsymbol{\mu}, \zeta \mathbb{I}_m/2)$  is a  $m$ -variate normal distribution having mean  $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_m]^T$  and  $\zeta \mathbb{I}_m/2$  variance-covariance matrix.

The Jeffreys prior of the  $\{\boldsymbol{\mu}, \zeta\}$  parameters is

$$\pi_J(\boldsymbol{\mu}, \zeta | \zeta_0 = 0) \propto 1/\sqrt{\zeta^{m+2}}, \quad (6.46)$$

where  $\mathcal{M}_0 = \{p(\bar{\mathbf{x}}, s^2 | \boldsymbol{\mu}, \zeta) : (\boldsymbol{\mu}, \zeta) \in \mathbb{R}^m \times \mathbb{R}^+\}$  is the data model and  $\zeta > \zeta_0 = 0$ . The reason of the explicit introduction of  $\zeta_0 = 0$  hyper-parameter will be clarified in the next section. The prior (6.46) results in the marginal likelihood [Wolfram Research Inc. (2012)]

$$\begin{aligned} Z(\bar{\mathbf{x}}, s^2 | \zeta_0 = 0) &= \int_0^\infty \int_{-\infty}^{+\infty} p(\bar{\mathbf{x}}, s^2 | \boldsymbol{\mu}, \zeta) \pi_J(\boldsymbol{\mu}, \zeta | \zeta_0 = 0) d\boldsymbol{\mu} d\zeta \\ &\propto \frac{\Gamma(m+1)}{\sqrt{(ms^2)^{m+2}} \Gamma(m/2)}, \end{aligned} \quad (6.47)$$

where the specification  $\zeta_0 = 0$  highlights the conditioning over the data-explaining model, and posterior distribution

$$p(\boldsymbol{\mu}, \zeta | \bar{\mathbf{x}}, s^2, \zeta_0 = 0) = \frac{m(ms^2)^m}{\sqrt{\pi^m} \Gamma(m+1)} \frac{\exp\left(-\frac{ms^2 + |\boldsymbol{\mu} - \bar{\mathbf{x}}|^2}{\zeta}\right)}{\sqrt{\zeta^{3m+2}}}. \quad (6.48)$$

After the means  $\boldsymbol{\mu}$  are integrated out, the posterior density and expected value of  $\zeta$  are

$$p(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0 = 0) = \int_{-\infty}^{+\infty} p(\boldsymbol{\mu}, \zeta | \bar{\mathbf{x}}, s^2, \zeta_0 = 0) d\boldsymbol{\mu} = \frac{(ms^2)^2 \exp(-ms^2/\zeta)}{\zeta^{m+1} \Gamma(m)} \quad (6.49)$$

and

$$\bar{\zeta} = \mathbb{E}(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0 = 0) = \int_0^\infty \zeta p(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0 = 0) d\zeta = \frac{ms^2}{m-1}, \quad (6.50a)$$

where  $m \geq 2$ , with a variance equal to

$$\text{Var}(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0 = 0) = \int_0^\infty \zeta^2 p(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0 = 0) d\zeta - \bar{\zeta}^2 = \frac{\bar{\zeta}^2}{m-2}, \quad (6.50b)$$

where  $m \geq 3$ .

Since  $2ms^2/\zeta$  is a  $\chi_m^2$  variable having  $m$  degrees of freedom,  $s^2$  is a frequentist biased estimator of  $\zeta$ . In fact,

$$\overline{s^2} = \mathbb{E}(s^2 | \zeta) = \frac{2m}{\zeta} \int_0^\infty s^2 \chi_m^2(2ms^2/\zeta) ds^2 = \zeta/2 \quad (6.51a)$$

and

$$\text{Var}(s^2 | \zeta) = \frac{2m}{\zeta} \int_0^\infty s^4 \chi_m^2(2ms^2/\zeta) ds^2 - (\overline{s^2})^2 = \zeta^2/m. \quad (6.51b)$$

Also in this case a paradox occurs. As  $m$  tends to the infinity, (6.50a) and (6.50b) predict that  $\zeta$  is certainly equal to  $s^2$ , but, (6.51a) and (6.51b) predict that  $\zeta$  is certainly equal to  $2s^2$ .

**Proposed solution.**

To explain the paradox, we observe that (6.46) gives for certain that  $\zeta$  is less than any positive number, no matter how small it is. Therefore, (6.46) encodes that  $\zeta$  is null, but we observe a non-zero pooled variance. To avoid this contradiction, the  $\zeta$ 's domain must be limited to  $[\zeta_0, \infty[$ , the hyper-parameter  $\zeta_0 = \sigma_0^2 > 0$  being unknown. Hence, the Jeffreys' distribution

$$\pi_J(\boldsymbol{\mu}, \zeta | \zeta_0) = \frac{m}{2} \sqrt{\frac{\zeta_0^m}{\zeta^{m+2}}} \vartheta(\zeta - \zeta_0), \quad (6.52)$$

where the Heaviside function  $\vartheta(z)$  is null for  $z < 0$  and one otherwise, substitutes for (6.46). This results in the marginal likelihood [Wolfram Research Inc. (2012)]

$$\begin{aligned} Z(\bar{\mathbf{x}}, s^2 | \zeta_0) &= \int_{\zeta_0}^{\infty} \int_{-\infty}^{+\infty} p(\bar{\mathbf{x}}, s^2 | \boldsymbol{\mu}, \zeta) \pi_J(\boldsymbol{\mu}, \zeta | \zeta_0) d\boldsymbol{\mu} d\zeta \\ &= \frac{m^2 \sqrt{\zeta_0^m} \Gamma(m, 0, ms^2/\zeta_0)}{4\sqrt{m^m} s^{m+2} \Gamma(m/2 + 1)} \end{aligned} \quad (6.53)$$

and posterior distribution

$$p(\boldsymbol{\mu}, \zeta | \bar{\mathbf{x}}, s^2, \zeta_0) = \frac{(ms^2)^m}{\sqrt{\pi^m} \Gamma(m, 0, ms^2/\zeta_0)} \frac{\exp\left(-\frac{ms^2 + |\boldsymbol{\mu} - \bar{\mathbf{x}}|^2}{\zeta}\right)}{\sqrt{\zeta^{3m+2}}}, \quad (6.54)$$

where  $\zeta > \zeta_0$ .

After integrating out the means  $\boldsymbol{\mu}$ , the posterior density and expected value of  $\zeta$  are

$$p(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0) = \int_{-\infty}^{+\infty} p(\boldsymbol{\mu}, \zeta | \bar{\mathbf{x}}, s^2, \zeta_0) d\boldsymbol{\mu} = \frac{(ms^2)^m \exp(-ms^2/\zeta)}{\zeta^{m+1} \Gamma(m, 0, ms^2/\zeta_0)}, \quad (6.55)$$

where  $\zeta > \zeta_0$ , and

$$E(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0) = \int_{\zeta_0}^{\infty} \zeta p(\zeta | \bar{\mathbf{x}}, s^2, \zeta_0) d\zeta = \frac{ms^2 \Gamma(m-1, 0, ms^2/\zeta_0)}{\Gamma(m, 0, ms^2/\zeta_0)}. \quad (6.56)$$

Different hyper-parameters, that is, different lower bound of the  $\zeta$  coordinate, correspond to different manifolds and to different models. In order to average over them, we need the Bayes theorem to calculate the model probability – consistently with the observed data – and, in turn, we need the prior of  $\zeta_0$ . By applying the Jeffreys rule to (6.53), which is the sampling distribution of  $\bar{\mathbf{x}}$  and  $s^2$  given  $\zeta_0$ , we obtain

$$\pi_J(\zeta_0) \propto 1/\zeta_0, \quad (6.57)$$

After combining (6.57) with (6.53), the posterior probability density of  $\zeta_0$  is

$$p(\zeta_0 | \bar{\mathbf{x}}, s^2) = \frac{m^2 \sqrt{\zeta_0^{m-2}} \Gamma(m, 0, ms^2/\zeta_0)}{4\sqrt{m^m} s^{m+2} \Gamma(m/2 + 1)}. \quad (6.58)$$

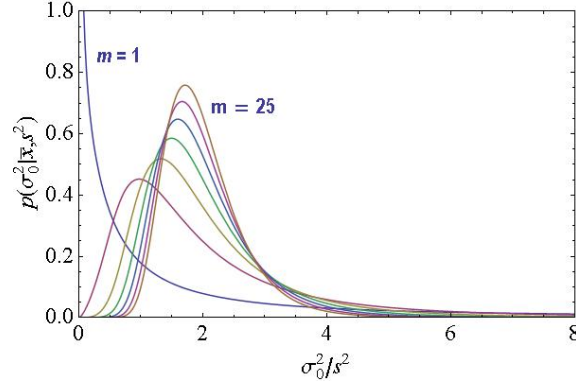


Figure 8: Probability density of the data models (6.45), where  $\zeta = \sigma^2 > \sigma_0^2$ , vs. the  $\sigma_0^2$  hyper-parameter. Different numbers of observed quantities, from  $m = 1$  to  $m = 25$ , are considered, in  $m = 4$  steps.

The odds on  $\zeta_0$  explaining the data are shown in Fig. 8. It is worth noting that, in the limit of many observed quantities, the lower bound of the model variance most supported by the data is twice the pooled variance.

The Neymann-Scott paradox is solved by observing that, after averaging (6.56) over  $\zeta_0$  with the (6.58) weights, the expected  $\zeta$  value is [Wolfram Research Inc. (2012)]

$$E(\zeta | \bar{x}, s^2) = \frac{2ms^2}{m-2}, \quad (6.59)$$

where  $m \geq 3$ . In the limit when  $m \rightarrow \infty$ ,  $E(\zeta | \bar{x}, s^2) = 2s^2$ , which is the frequentist unbiased estimate.

## 6.6 marginalization paradox

The post-data distribution (6.49) of  $\zeta = \sigma^2$  depends only on the pooled variance  $s^2$ . Since  $s^2$  and  $\bar{x}$  are independent chi-square and normal variables, respectively, it might seem that the  $\bar{x}$  data are irrelevant and can be omitted [Dawid et al. (1973); Bernardo (1979); Kass and Wasserman (1996)]. But, it is not so.

To investigate the contribution of  $\bar{x}$  to inferences about  $\zeta$ , let us discard it and start from the  $\mathcal{X} = \{p(s^2 | \zeta, m) : \zeta \in \mathbb{R}^+\}$  model, where the sampling distribution is

$$p(s^2 | \zeta, m) = \frac{2m}{\zeta} \chi_m^2(2ms^2/\zeta). \quad (6.60)$$

The Jeffreys prior of  $\zeta$  is  $\pi_J(\zeta | \mathcal{X}) \propto 1/\zeta$ . This results in the marginal likelihood

$$Z(s^2 | m, \mathcal{X}) = \int_0^\infty p(s^2 | \zeta, m, \mathcal{X}) \pi_J(\zeta | \mathcal{X}) d\zeta \propto 1/s^2, \quad (6.61)$$

posterior-distribution

$$p(\zeta|s^2, m, \mathcal{X}) = \frac{\sqrt{(ms^2)^m} \exp(-ms^2/\zeta)}{\sqrt{\zeta^{m+2}} \Gamma(m/2)}, \quad (6.62)$$

and expectation

$$\bar{\zeta} = \mathbb{E}(\zeta|s^2, m, \mathcal{X}) = \int_0^\infty \zeta p(\zeta|s^2, m, \mathcal{X}) d\zeta = \frac{2ms^2}{m-2} \quad (6.63)$$

that are different from (6.49) and (6.50a).

To explain the paradox, we observe that – as shown by the joint posterior distribution (6.48) –  $\mu$  and  $\zeta$  are not independent [Jaynes (2012)]. Therefore,  $\mu$  is relevant and its posterior distribution affect (6.49). The posterior distribution (6.62) differs from (6.49) because the information delivered by  $\bar{\mathbf{x}}$  has been neglected. As a consequence the  $\zeta$  variance conditioned to  $\mathcal{X}$ ,

$$\text{Var}(\zeta|s^2, m, \mathcal{X}) = \int_0^\infty \zeta^2 p(\zeta|s^2, m, \mathcal{X}) d\zeta - \bar{\zeta}^2 = \frac{2\bar{\zeta}^2}{m-4}, \quad (6.64)$$

is bigger than (6.50b). It is also worth noting that (6.63) does not suffer from the (6.50a) inconsistency and it is the same as (6.59). Our conjectured explanation is as follows. Differently from (6.46), which is conditioned to  $\mathcal{M}_0$ , the  $\pi_J(\zeta|\mathcal{X}) \propto 1/\zeta$  distribution assigns the same probability to any  $\zeta$ 's sub-domain including the zero or infinity and a null probability to any sub-domain including none of the two. Therefore, the biasing effect of (6.46) is removed.

## 7 Conclusions

In metrology, assessing the measurement uncertainty requires invariant metrics of the data models and covariant priors. In short,

- the posterior probability density must be covariant for model re-parametrization and, consequently, the prior density must be covariant;
- the Jeffreys priors, which are derived from the volume element of model manifold equipped with the information metric, are sensible choices;
- the posterior probability of a selected model must not be null; therefore, improper priors are not allowed;
- to avoid paradox and inconsistencies, the computation of the posterior probability may involve hierarchical models, model selection, and averaging.

These ideas were tested by the application to a number of key paradoxical problems. Non-uniquenesses were identified in the data-explaining models. In addition, parametric



models having an infinite volume, corresponding to improper priors, were found to hide unacknowledged information. After recognizing these issues and taking them into account via hierarchical modelling and model averaging, inconsistencies and paradoxes disappeared.

## References

- Amari, S., Nagaoka, H., and Harada, D. (2007). *Methods of Information Geometry*. Translations of Mathematical Monographs vol. 191. Oxford: Oxford University Press. [2](#), [4](#)
- Arwini, K. and Dodson, C. (2008). *Information Geometry: Near Randomness and Near Independence*. Lecture Notes in Mathematics 1953. Berlin Heidelberg: Springer. [2](#), [4](#)
- Attivissimo, F., Giaquinto, N., and Savino, M. (2012). “A Bayesian paradox and its impact on the GUM approach to uncertainty.” *Measurement*, 45(9): 2194–2202. [2](#), [13](#)
- Berger, J. O. and Bernardo, J. M. (1992a). “On the development of reference priors.” In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics: Proceedings of the Fourth Valencia International Meeting*. Oxford: Clarendon Press. [2](#)
- (1992b). “Ordered group reference priors with application to the multinomial problem.” *Biometrika*, 79: 25–37. [2](#)
- Berger, J. O., Bernardo, J. M., and Sun, D. (2009). “The formal definition of reference priors.” *Ann. Statist.*, 37(2): 905–938. [2](#)
- (2015). “Overall Objective Priors.” *Bayesian Anal.*, 10(1): 189–221. [2](#), [8](#), [12](#), [13](#)
- Berger, J. O. and Sun, D. (2008). “Objective priors for the bivariate normal model.” *Ann. Statist.*, 36(2): 963–982. [2](#)
- Bernardo, J. M. (1979). “Reference Posterior Distributions for Bayesian Inference.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2): 113–147. [2](#), [8](#), [13](#), [23](#)
- (1989). “The Geometry of Asymptotic Inference: Comment: On Multivariate Jeffreys’ Priors.” *Statistical Science*, 4(3): 227–229. [2](#), [12](#), [13](#)
- (2005). “Reference analysis.” In K., D. D. and R., R. C. (eds.), *Handbook of Statistics* 25. Amsterdam: North Holland. [2](#), [13](#)
- Bodnar, O., Link, A., and Elster, C. (2015). “Objective Bayesian Inference for a Generalized Marginal Random Effects Model.” *Bayesian Anal.*, advance publication. [2](#)
- Carobbi, C. (2014). “Bayesian inference on a squared quantity.” *Measurement*, 48(1): 13–20. [2](#), [13](#)
- Clyde, M. and George, E. I. (2004). “Model Uncertainty.” *Statistical Science*, 19(1): 81–94. [7](#)

- Costa, S. I., Santos, S. A., and Strapasson, J. E. (2014). “Fisher information distance: A geometrical reading.” *Discrete Applied Mathematics*, 197: 59–69. [2](#)
- D’Agostini, G. (2003). *Bayesian Reasoning in Data Analysis: A Critical Introduction*. Singapore: World Scientific. [1](#), [2](#)
- Datta, G. S. and Ghosh, M. (1995). “Some Remarks on Noninformative Priors.” *Journal of the American Statistical Association*, 90(432): 1357–1363. [2](#)
- (1996). “On the invariance of noninformative priors.” *Ann. Statist.*, 24(1): 141–159. [2](#)
- Dawid, A. P., Stone, M., and Zidek, J. V. (1973). “Marginalization Paradoxes in Bayesian and Structural Inference.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(2): pp. 189–233. [23](#)
- Dose, V. (2007). “Bayesian estimate of the Newtonian constant of gravitation.” *Measurement Science and Technology*, 18(1): 176–182. [1](#)
- Elster, C. and Toman, B. (2010). “Analysis of key comparisons: estimating laboratories’ biases by a fixed effects model using Bayesian model averaging.” *Metrologia*, 47(3): 113–119. [1](#)
- Ghosh, M. (2011). “Objective Priors: An Introduction for Frequentists.” *Statist. Sci.*, 26(2): 187–202. [2](#)
- Jaynes, E. and Bretthorst, G. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press. [1](#), [2](#)
- Jaynes, E. T. (2012). “Marginalization and Prior Probabilities.” In Rosenkrantz, R. D. (ed.), *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. Dordrecht: Springer. [24](#)
- Jeffreys, H. (1946). “An Invariant Form for the Prior Probability in Estimation Problems.” *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186: 453–461. [2](#), [5](#), [10](#)
- (1998). *The Theory of Probability*. Oxford: Oxford University Press. [2](#), [5](#), [10](#)
- Kass, R. E. (1989). “The Geometry of Asymptotic Inference.” *Statist. Sci.*, 4(3): 188–219. [2](#)
- Kass, R. E. and Wasserman, L. (1996). “The Selection of Prior Distributions by Formal Rules.” *Journal of the American Statistical Association*, 91(435): 1343–1370. [2](#), [23](#)
- MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press. [1](#), [2](#), [7](#)
- Mana, G. (2015). “Model uncertainty and reference value of the Planck constant.” *Measurement*, submitted. [1](#)
- Mana, G., Giuliano Albo, P. A., and Lago, S. (2014). “Bayesian estimate of the degree of a polynomial given a noisy data sample.” *Measurement*, 55: 564–570. [1](#)

- Mana, G., Massa, E., and Predescu, M. (2012). “Model selection in the average of inconsistent data: an analysis of the measured Planck-constant values.” *Metrologia*, 49(4): 492–500. [1](#)
- Neyman, J. and Scott, E. L. (1948). “Consistent Estimates Based on Partially Consistent Observations.” *Econometrica*, 16(1): 1–32. [2](#), [20](#)
- Rao, C. R. (1945). “Information and the accuracy attainable in the estimation of statistical parameters.” *Bull. Calcutta Math. Soc.*, 37: 81–89. [2](#)
- Samworth, R. J. (2012). “Steins Paradox.” *Eureka*, 62: 38–41. [2](#), [13](#)
- Sivia, D. and Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial*. Oxford: Oxford University Press. [1](#), [2](#), [7](#)
- Stein, C. (1959). “An Example of Wide Discrepancy Between Fiducial and Confidence Intervals.” *Ann. Math. Statist.*, 30(4): 877–880. [2](#), [13](#)
- Toman, B., Fischer, J., and Elster, C. (2012). “Alternative analyses of measurements of the Planck constant.” *Metrologia*, 49(4): 567–571. [1](#)
- von der Linden, W., Dose, V., and von Toussaint, U. (2014). *Bayesian Probability Theory: Applications in the Physical Sciences*. Cambridge: Cambridge University Press. [1](#), [2](#), [5](#), [7](#)
- Wolfram Research Inc. (2012). *Mathematica*. Version 9.0. Champaign, Illinois: Wolfram Research, Inc. [7](#), [8](#), [11](#), [13](#), [16](#), [18](#), [19](#), [21](#), [22](#), [23](#)

### Acknowledgments

This work was jointly funded by the European Metrology Research Programme (EMRP) participating countries within the European Association of National Metrology Institutes (EURAMET), the European Union, and the Italian ministry of education, university, and research (awarded project P6-2013, implementation of the new SI).